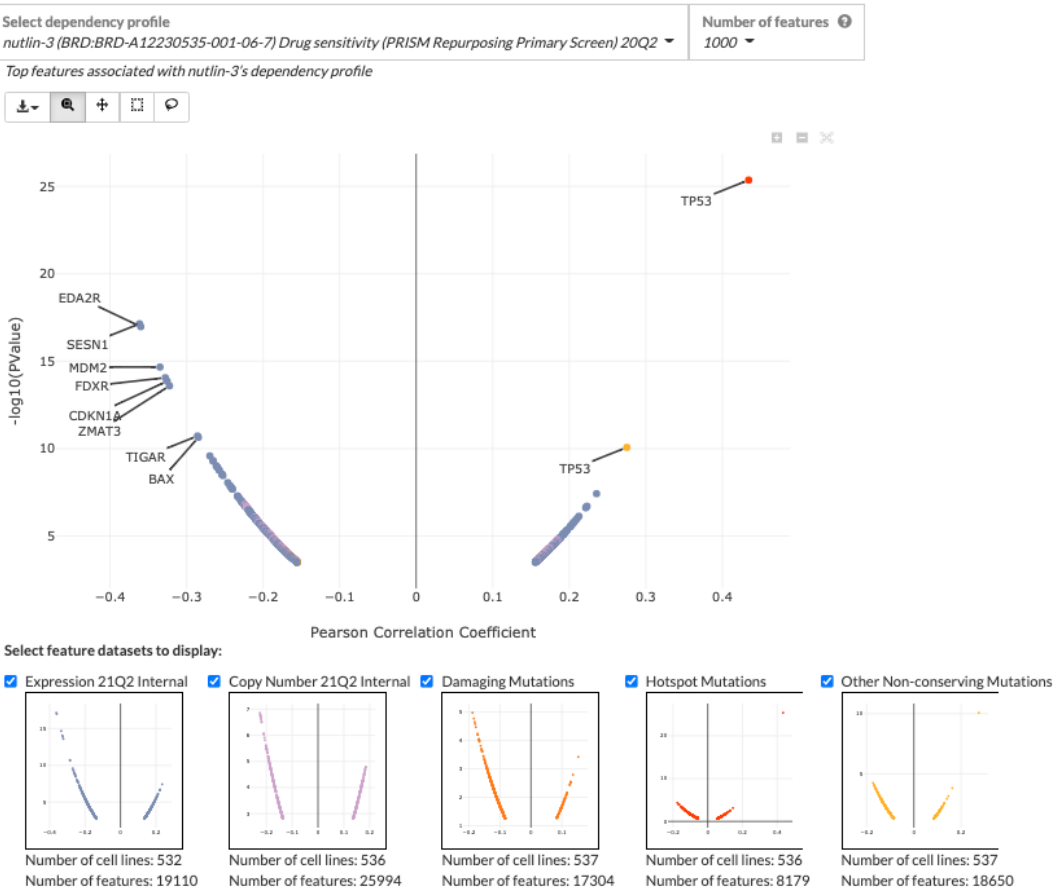
Genomic Associations

How to use the genomic associations tab

This tab provides information about the baseline genomic features that correlate with a given gene dependency or compound sensitivity profile. This information can be helpful for understanding the mechanism of therapeutic targets and compounds. This tab differs from the predictability tab because it is designed to support the exploration and interpretation of many intercorrelated features.

Here is how the genomic associations tab can be used to rediscover the relationship between sensitivity to the MDM2 inhibitor Nutlin-3 and p53 status.

1. The volcano plot shows the top features associated with Nutlin-3 sensitivity with color indicating feature type. TP53 mutations are associated with decreased Nutlin-3 sensitivity (higher AUC), and the expression of a number of genes, including EDA2R, SESN1, and MDM2, is associated with increased Nutlin-3 sensitivity (lower AUC).

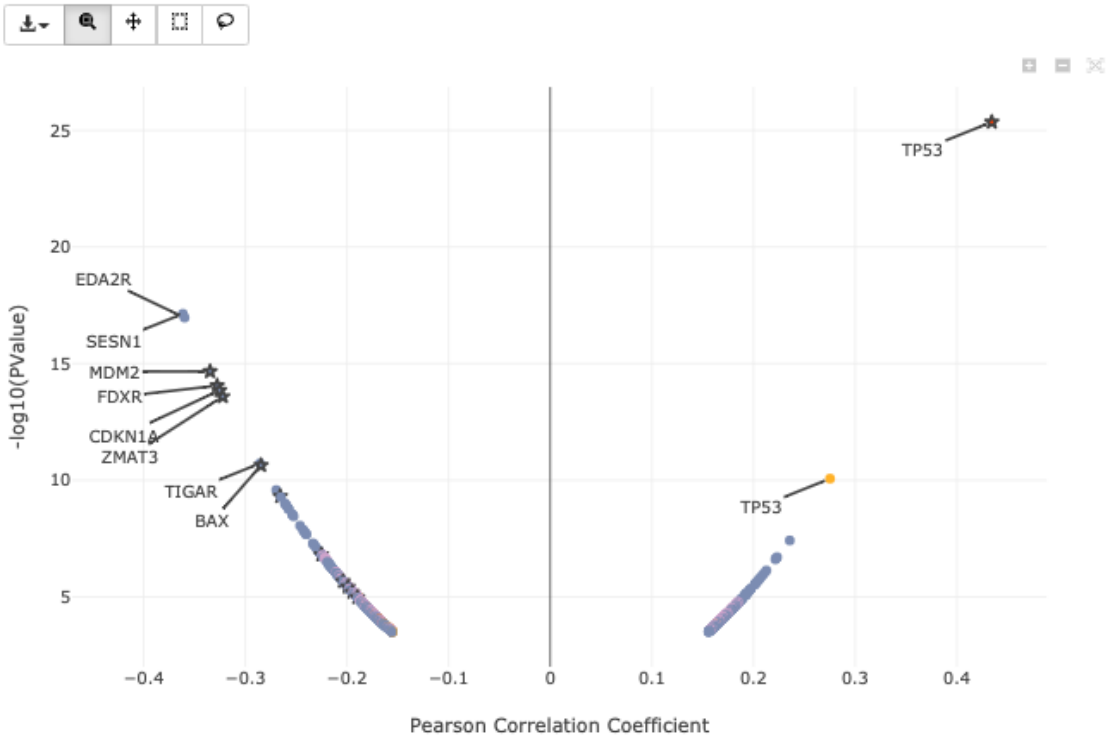


2. Gene set overrepresentation is calculated using a subset of the top genes associated with Nutlin-3 sensitivity. The default is the 200 genes with the highest absolute correlation. The most significant gene set is the Hallmark p53 pathway which is enriched in the negatively correlated features.

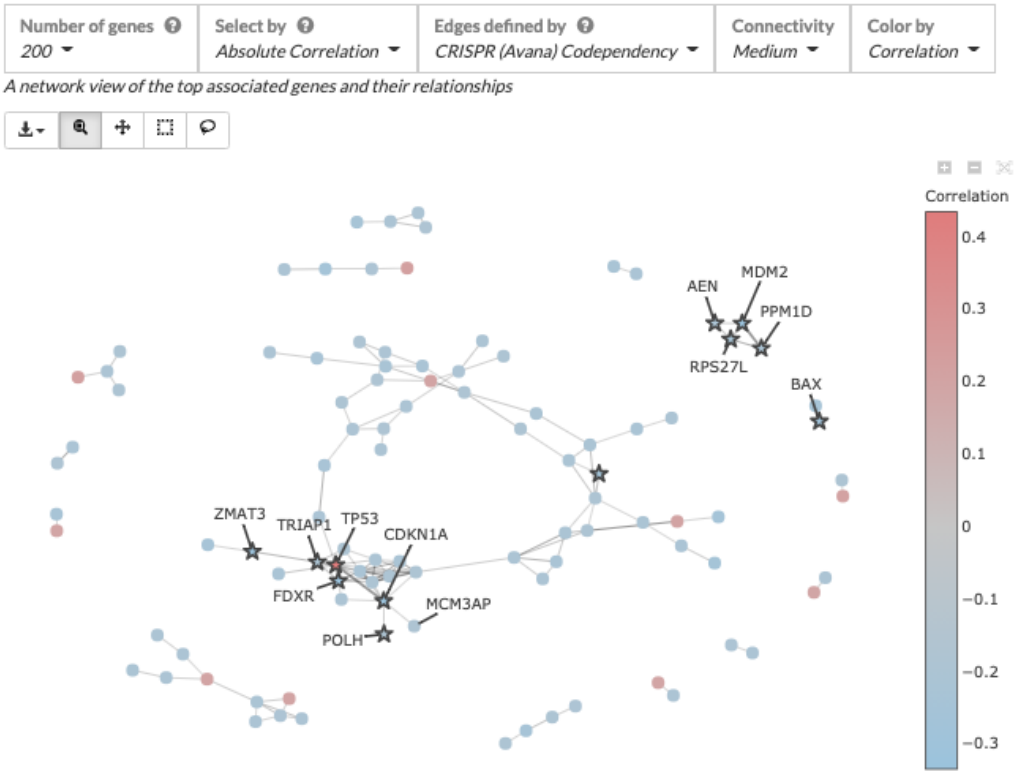
Gene set overrepresentation of selected positively and negatively associated genes

Highlight	Cor Sign	Geneset	-log10(P-Value)	Set size
			0.49 – 9.2	15 – 3.6e+2
Highlight	Neg	HALLMARK P53 PATHWAY	9.17	200
Highlight	Neg	PID P53 DOWNSTREAM PATHWAY	6.02	137
Highlight	Neg	KEGG P53 SIGNALING PATHWAY	5.64	68
Highlight	Neg	PID TAP63 PATHWAY	3.91	54
Highlight	Pos	HALLMARK INTERFERON ALPHA RESPONSE	1.46	97
Highlight	Neg	REACTOME TRANSCRIPTIONAL REGULATION BY TP53	1.18	363
Highlight	Neg	BIOCARTA P53 PATHWAY	1.02	16
Highlight	Neg	PID P73PATHWAY	0.874	79
		REACTOME TP53 REGULATES TRANSCRIPTION OF GENES		

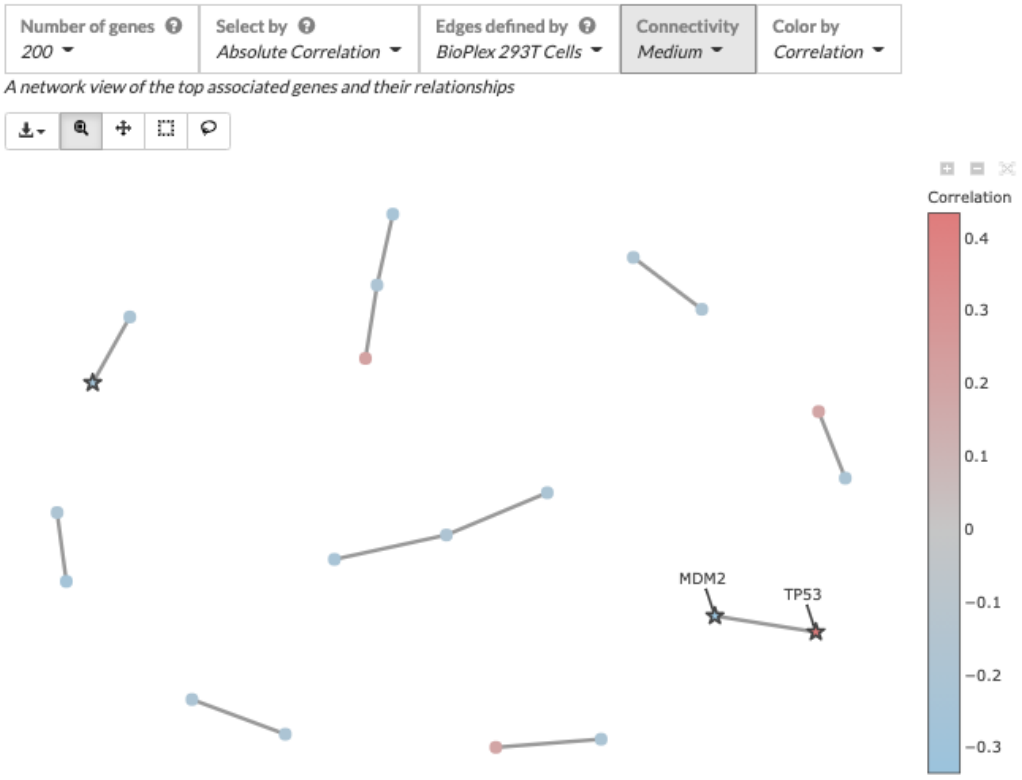
3. Clicking the highlight button next to the Hallmark p53 pathway stars the genes in this set, which reveals that many of the most significant features associated with Nutin-3 sensitivity are p53 pathway genes such as MDM2, FDXR, and CDKN1A.



4. The network shows relationships between the same subset of the top genes. The default network uses codependency data from DepMap. This type of network has been shown to cluster genes into functional modules that often reflect known protein complexes (Pan et al., 2018). The p53 pathway genes form two major clusters, one containing TP53 and the other containing the negative regulators for p53 MDM2 and PPM1D.



5. Changing the network edges to protein-protein interactions from the BioPlex 293T database highlights the known physical interaction between MDM2 and TP53.



Methodology

Genomic features

The latest versions of these datasets are tested for association with the dependency of interest:

1. Expression
2. Copy Number
3. Damaging Mutations
4. Hotspot Mutations
5. Other Non-conserving Mutations

Association test

The Pearson correlation and corresponding p-values are calculated for each genomic feature.

Top associated genes

The top N (set using dropdown) associated genes are used to calculate gene set overrepresentation and create a network. There are four options for selecting these genes:

Absolute correlation: N genes with the largest absolute correlation.
Max correlation: N genes with the largest positive correlation.
Min correlation: N genes with the largest negative correlation.
-log10(P): N genes with the largest -log10(p_value).

The network usually has fewer than N genes because only genes connected to other genes are included. The top N genes are still used for overrepresentation analysis.

Edge types

A number of different datasets can be used to define edges.

CRISPR (Avana) Codependency
CRISPR codependency is calculated based on the DepMap 20Q2 achilles_gene_effect file. Edge weights are the Pearson correlation coefficient between gene pairs.

MSigDB curated pathways

An edge connecting two genes has a higher weight the more similar the two genes are to each other. Intuitively, two genes are similar if they are each members of the same, or similar, curated gene sets. Specifically, each gene is represented by a 64-dimensional vector learned from the gene set membership data of MSigDB's C2 Canonical Pathways (v7.0) and Hallmark (v7.1) gene set collections. StarSpace was used to learn the representations. Cosine distance was used to define the similarity between two gene representations.

STRING DB
Three separate STRING DB edge types are used: protein-protein interaction (PPi), literature, and combined. Edge weights come directly from the String-DB Homo sapiens v11 protein.links.detailed file.

Feature Correlation
Feature Correlation is calculated based on the DepMap 21Q2 expression, copy number, and mutation data. Edge weights are the Pearson correlation coefficient between feature pairs.

BioPlex PPi 293T Cells
Bioplex 3.0 interactions in 293T cells from BioPlex Interactions.

Connectivity

The connectivity setting specifies the size of the edge database used to connect the top N genes. High, medium, and low connectivity mean the same thing for all of the edge types:

> **High:** top 1% by weight of possible gene-gene edges are used.
> **Medium:** top 0.2% by weight of possible gene-gene edges are used.
> **Low:** top 0.004% by weight of possible gene-gene edges are used.

The number of possible gene-gene edges is set at 19321^2 since there are 19321 named protein-coding genes in the HGNC database. Since connectivity is set globally, the number of edges will vary between gene lists, and it is possible that some gene lists will have no edges.

Network layout

The x and y position of each gene in the network is calculated using the Fruchterman-Reingold force-directed layout algorithm. Specifically, an undirected graph with weighted edges specified by the selected edge type and connectivity is input into the *layout_with_fr* function from the igraph package. Genes that are not connected to any other genes are not included in the network.

Gene set overrepresentation

Gene set overrepresentation is calculated separately for genes with positive and negative correlations within the top N genes. P-values for overrepresentation in the MSigDB v7.1 hallmark (H) and curated pathway (C2) gene sets are calculated with Fisher's exact test.